

Community detection by label propagation with compression of flow

Jihui Han^a, Wei Li^b, Zhu Su, Longfeng Zhao, and Weibing Deng

Complexity Science Center, Institute of Particle Physics, Central China Normal University, Wuhan 430079, China

the date of receipt and acceptance should be inserted later

Abstract. The label propagation algorithm (LPA) has been proved to be a fast and effective method for detecting communities in large complex networks. However, its performance is subject to the non-stable and trivial solutions of the problem. In this paper, we propose a modified label propagation algorithm LPAf to efficiently detect community structures in networks. Instead of the majority voting rule of the basic LPA, LPAf updates the label of a node by considering the compression of a description of random walks on a network. A multi-step greedy agglomerative strategy is employed to enable LPAf to escape the local optimum. Furthermore, an incomplete update condition is also adopted to speed up the convergence. Experimental results on both synthetic and real-world networks confirm the effectiveness of our algorithm.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Hc Networks and genealogical trees

1 Introduction

Real-life complex systems in many research fields such as biology, sociology, economy and computer science, can be studied as networks with nodes representing for individuals and links for interactions or relations between individuals. Many networks exhibit the so-called community structure: nodes tend to organize themselves in groups such that connections are denser within groups while sparser between groups. Community structure is a prominent feature of complex networks, as it often represents functional modules with nodes of common properties and accounts for the functionality of the system. Community detection

enables us to probe the organization and functional behavior of real-world systems, therefore has been paid much attention and applied to many kinds of networks, including the collaboration networks [1], social networks [2], and biological networks [3], etc.

Community detection has been studied as the graph partitioning in computer science for decades and remains quite challenging. Algorithms to detect reasonably good quality communities have been proposed and improved extensively [4], especially in recent years, such as Girvan-Newman algorithm [5], spectral clustering [6, 7], multi-state spin model [8–10] (e.g., q-state Potts model), random walk [11–13], modularity optimization [14–17] and statistical inference [18–21].

^a e-mail: jh@mails.ccnu.edu.cn

^b e-mail: liw@mail.ccnu.edu.cn

As one of the fastest algorithms for community detection, the label propagation algorithm (LPA) [22] uses the network structure alone to guide its process and requires neither parameters nor optimization of any object function. It starts by assigning each node a unique label, indicating the community it belongs to. At every label propagation step, each node sequentially updates its label to a new one that most of its neighbors own. If more than one label is the most frequent, the new label is chosen randomly among them. The label propagation step is performed iteratively until each node has a label that is the most frequent among its neighbors'. Through this iterative process, the densely connected groups of nodes form consensus on one label to form communities. Finally, LPA converges when no node changes its label anymore. Therefore, nodes with the same label are classified into the same community. In addition to its nearly linear time complexity, LPA introduces no parameter and requires no priori information of communities, and thus is suitable to process large-scale networks with millions of nodes and edges.

Due to the frequent tie-breaks and the random order update strategy, LPA usually delivers multiple partitions starting from the same initial condition, with different random seeds. Raghavan et al. [22] proposed to label each node with the set of all labels it has in different partitions to detect possible overlapping communities. However, in a recent paper, Tibely and Kertesz [23] showed that this method was equivalent to finding the local minima in a simple zero-temperature kinetic Potts model. The number of such local minima was found to be much larger than the number of nodes in the underlying network. Aggregating partitions suggested by Raghavan et al. [22] leads to a fragmentation of the resulting partition in small clusters when the number of aggregated partitions is large.

In order to eliminate undesired solutions, Barber and Clark [24] proposed a modularity-specialized LPA (LPAm)

to constrain the label propagation process, which is inclined to get stuck in poor local maximum of modularity. To solve this problem, Liu et al. [25] introduced an advanced modularity-specialized LPA (LPAm+), which is more stable than LPAm. Due to the usage of modularity, the capability of both algorithms will be affected by the resolution limit [26].

Leung et al. [27] have found that LPA often yields partitions with one giant community together with much smaller ones when applied to online social networks. In order to avoid such a disturbing feature, they proposed a modified method by adding a decreasing score assignment for each label in label propagation process (LPA- δ), which encourages the formation of a stronger local community and deters the occurrence of trivial solutions. Tests of LPA- δ on the LFR benchmark produced good results [28]. To save the running time of LPA- δ , Leung et al. proposed to avoid label update of those nodes with high neighbor purity [27]. Since the neighbor purity ignores contribution of the small degree nodes to the community detection, the detection precision is not high enough.

In this paper, we propose the LPAf which introduces a new update rule to update the label of a node by taking into account the compression of flow (random walks on a network), and uses an incomplete update condition in label propagation process to speed up the convergence. Like LPAm+, LPAf employs a multi-step greedy agglomerative algorithm (MSG) [29] to simultaneously merge multiple pairs of communities. Although LPAf is also applicable to weighted and directed networks, we currently focus on unweighted and undirected networks. The paper is organized as follows. In Sec. 2, we present our new method in detail. Experimental results on synthetic and real-world networks are shown in Sec. 3. Finally, the main findings are summarized in Sec. 4.

2 Algorithm

To reveal community structures in networks, Rosvall and Bergstrom [30] introduced an information theoretic approach (known as Infomap algorithm). They use the probability flow of random walks on a network as a proxy for information flows in real systems and decompose the network into communities by compressing a description of the probability flow.

For a network partition C of n nodes containing c communities, the average description length of random walks is defined as [30],

$$L(C) = q_{\sim} H(\Omega) + \sum_{i=1}^c p_{i\sim} H(P^i), \quad (1)$$

where

$$H(\Omega) = - \sum_{i=1}^c \frac{q_{i\sim}}{q_{\sim}} \log \left(\frac{q_{i\sim}}{q_{\sim}} \right), \quad (2)$$

and

$$H(P^i) = - \frac{q_{i\sim}}{p_{i\sim}} \log \left(\frac{q_{i\sim}}{p_{i\sim}} \right) - \sum_{\alpha \in i} \frac{p_{\alpha}}{p_{i\sim}} \log \left(\frac{p_{\alpha}}{p_{i\sim}} \right), \quad (3)$$

in which $\alpha = 1, 2, \dots, n$ and $i = 1, 2, \dots, c$.

Here $q_{i\sim}$ is the probability of exiting community i , $q_{\sim} = \sum_{i=1}^c q_{i\sim}$ is the probability that the random walker switches to a different community at any given time step, p_{α} is the probability of visiting node α and $p_{i\sim} = \sum_{\alpha \in i} p_{\alpha} + q_{i\sim}$ is the fraction of time the random walker spends in community i plus the probability of exiting that community.

By combining Eqs. (1), (2) and (3), the expanded form of map equation can be written as,

$$L(C) = q_{\sim} \log(q_{\sim}) - 2 \sum_{i=1}^c q_{i\sim} \log(q_{i\sim}) - \sum_{\alpha=1}^n p_{\alpha} \log(p_{\alpha}) + \sum_{i=1}^c p_{i\sim} \log(p_{i\sim}). \quad (4)$$

Note that the term $\sum_1^n p_{\alpha} \log p_{\alpha}$ is independent of partitioning. Consequently, when we update the label of node α from i to j , it is sufficient to only keep track of changes of $q_{i\sim}$ and $q_{j\sim}$. They can be easily derived for any update

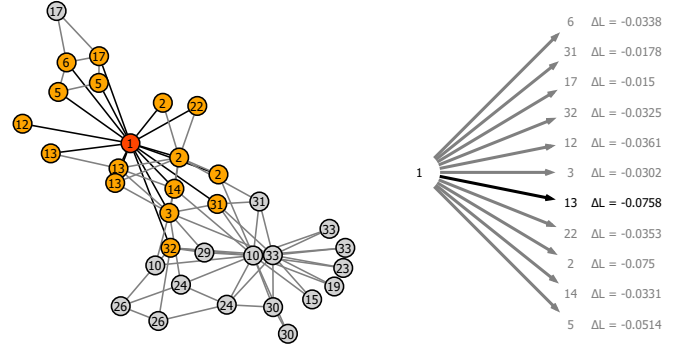


Fig. 1. (Color online) Snapshot of a label propagation step. Labels represent communities that nodes belong to. The node to be updated and its neighbors are orange and light orange respectively. Changes in average description length are shown on the right panel. The minimum of ΔL is highlighted by dark gray. Thus, according to our new update rule, the node ought to change its label from 1 to 13 in this case.

event, and updating them is fast and straightforward (see Appendix A for details).

We extend the LPA by modifying the label update rule so that the average description length $L(C)$ can be minimized. When update the label for α , we pick the one with the smallest ΔL (as illustrated on *karate* network in Fig. 1). Hence, our new update rule can be expressed as,

$$l_{\alpha}^{new} = \arg \min_{l \in N_l(\alpha)} \Delta L(\alpha, l_{\alpha}, l), \quad (5)$$

where l_{α} is the current label for node α , l_{α}^{new} is the new label for node α , $N_l(\alpha)$ includes the labels of the neighboring nodes of α , $\Delta L(\alpha, i, j)$ is the change of L when update the label of node α from community i to j (see Appendix A for details), and $\arg \min_l$ returns the label l that minimizes ΔL . If more than one label shares the same minimum of ΔL , the new label is chosen randomly among them. The label propagation step is performed iteratively until L no longer decreases.

In our tests, this update rule helps form local subgroups. However, it alone does not provide satisfying performance in dealing with large-scale networks, as it usually gets stuck in poor local minima in L space.

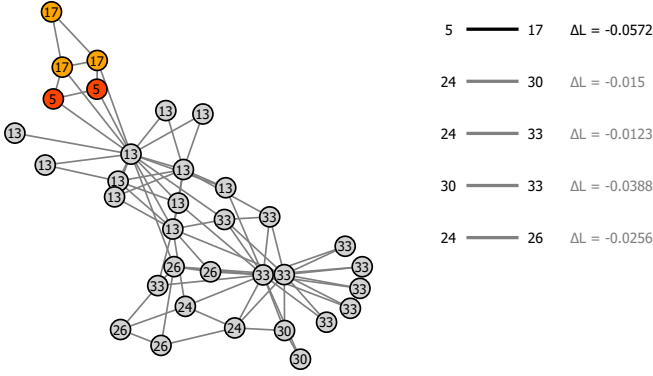


Fig. 2. (Color online) Snapshot of a merging event. Labels represent communities that nodes belong to. Changes in average description length for merging pairs of communities are shown on the right panel. In this case, communities 5 and 17 should be merged firstly as the merging of them leads to the smallest ΔL , followed by 30 and 33, 24 and 26. Community pairs 24 and 30, 24 and 33 are excluded because community 24 has already been merged before.

In order to escape the local minimum, we adopted a greedy rule for merging communities that minimizes L , i.e., when the LPA with our new update rule gets stuck in a local minimum (no decrease in L can be achieved via further label propagation), we calculate the changes of L for merging pairs of communities, and merge those pairs that decrease L the most. In actual operation, we employ the MSG technique to simultaneously merge multiple pairs of communities (as illustrated in Fig. 2). After merging communities, we escape the local minimum. Then we should perform another round of label propagation using the new update rule. This is analogous to downhill into another local minimum. However, it is not guaranteed that the new local minimum reached is good enough. Hence the above process should be repeated indefinitely until L no longer decreases.

To avoid unnecessary updates in each iteration of LPAf, the incomplete update condition proposed in Ref. [31] was adopted. Consequently, we only update the labels of the active nodes which would change their labels if they attempt to update. A list containing all currently active

nodes is maintained to allow the algorithm to finish execution when the list is empty (i.e., we only track the nodes that potentially change their labels). The pseudo-code of our algorithm is presented in Algorithm 1.

Algorithm 1 LPAf

- 1: each node is assigned a unique label
 - 2: perform label propagation using our new update rule
 - 3: **while** \exists community pairs with $\Delta L < 0$ **do**
 - 4: merge those community pairs using the MSG;
 - 5: perform label propagation using our new update rule;
 - 6: **end while**
-

3 Results

Many metrics have been proposed to quantify the quality of a network partition. When the ground truth is unknown, a common measure for the significance of the identified community structure is *modularity* Q [5], which is defined as,

$$Q = \frac{1}{2m} \sum_{u,v=1}^n (A_{uv} - P_{uv}) \delta(l_u, l_v), \quad (6)$$

where m is the total number of edges in the network. $A_{uv} = 1$ if nodes u and v are connected and 0 otherwise, $P_{uv} = k_u k_v / 2m$ is the probability in the null model that an edge exists between nodes u and v , and $\delta(i, j)$ is the Kronecker function: two vertices u and v provide a non-zero contribution to the value of Q if and only if they belong to the same community. The concept of *modularity* is based on the idea that a random graph is not expected to exhibit the community structure.

For a more sufficient assessment of the significance of detected communities, we also adopt the *modularity density* Q_{ds} [32] and the *conductance* Φ [33] metrics.

Given an undirected network, the modularity density is defined as

$$Q_{ds} = \sum_{c_i \in C} \left[\frac{|E_{c_i}^{in}|}{m} d_{c_i} - \left(\frac{2|E_{c_i}^{in}| + |E_{c_i}^{out}|}{2m} \right)^2 - \sum_{c_j \in C, c_j \neq c_i} \frac{|E_{c_i, c_j}|}{2m} d_{c_i, c_j} \right], \quad (7)$$

where C is the set of all the communities, c_i is any given community in C , $d_{c_i} = \frac{2|E_{c_i}^{in}|}{|c_i|(|c_i|-1)}$ is the internal density of community c_i , $d_{c_i, c_j} = \frac{|E_{c_i, c_j}|}{|c_i||c_j|}$ is the pair-wise density between communities c_i and c_j , $|E_{c_i}^{in}|$ is the number of edges between nodes within community c_i , $|E_{c_i}^{out}|$ is the number of edges from the nodes in community c_i to the nodes of other communities, and $|E_{c_i, c_j}|$ is the number of edges between communities c_i and c_j . Compared to modularity, the *modularity density* is an improved measurement for assessing the quality of communities, since it does not suffer from the well-known resolution limit of modularity.

For a community c_i , the conductance is defined as

$$\Phi(c_i) = \frac{\sum_{u \in c_i, v \notin c_i} A_{uv}}{\sum_{u \in c_i} k_u}, \quad (8)$$

where k_u is the degree of node u . Informally, *conductance* is the fraction of total edge volume that points outside the community c_i . Lower values of *conductance* imply that the communities have more internal connections than external ones, and thus represent more significant communities. Due to the fact that conductance cannot be easily extended to an entire community structure of a network, results are commonly assessed at different scales separately in the form of *network community profile (NCP)* [34] plots.

For networks with known community structures, two metrics from the field of information theory [35] are adopted to compare identified communities with the true ones. The first one, *normalized mutual information (NMI)* [36], estimates the amount of information correctly extracted by the detection algorithms and has become a de facto standard to quantify the quality of a detected partition with

respect to the ground truth. It is defined as,

$$NMI(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}, \quad (9)$$

where X and Y denote two partitions of the network, $I(X, Y) = H(X) - H(X|Y)$, $H(X)$ is the Shannon entropy of X and $H(X|Y)$ is the conditional entropy of X given Y . NMI equals 1 if the detected partition is identical to the real one, whereas it has an expected value of 0 if the detected partition is totally independent of the real one.

The second metric is the *variation of information (VOI)* [37], which has several desirable properties with respect to NMI . Specifically, it can be regarded as a kind of distance in the space of partitions. VOI of X and Y is defined as

$$VOI(X, Y) = H(X|Y) + H(Y|X). \quad (10)$$

Thus, lower values represent higher similarities between partitions. The value of VOI ranges from 0 to $\log N$, where N is the network size. Therefore, we divide the obtained values by $\log N$ for meaningful comparisons.

We have tested our algorithm on both synthetic and real-world networks. For comparisons, five algorithms, the original LPA [22], the neighbor strength driven LPA (nsdLPA) [31], the Louvain method [14], the Infomap algorithm [30], and the fine-tuned modularity density algorithm (FineTune) [38], are included in the experiments as references. The nsdLPA enhances the basic LPA by taking into account the positive neighborhood strength, and is generally efficient in practice [31]. The Louvain method is a greedy optimization algorithm that attempts to optimize the modularity of a partition, which usually produces high modularity values and is by far one of the most widely used method for detecting communities in large networks [14]. The Infomap algorithm decomposes a network into communities by compressing a description of information flow on the network as mentioned above [30]. The FineTune algorithm iteratively attempts to improve the mod-

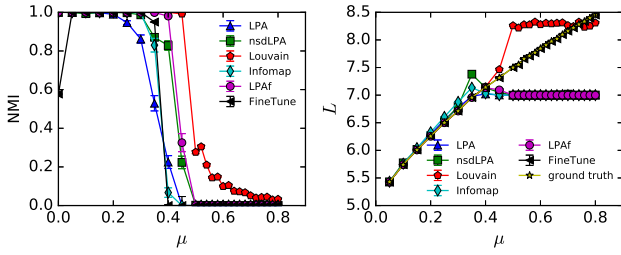


Fig. 3. (Color online) Average NMI and L for different algorithms as a function of μ on the GN benchmark networks. Each data point is computed over 100 different network realizations.

ularity density measurement by splitting and merging the given network community structure [38].

3.1 Tests on synthetic networks

We first tested our method on the well-known GN benchmark [39], and compared the results to the counterparts of other methods. The GN benchmark network consists of 128 nodes, each with expected degree 16, which are divided into four groups with 32 nodes each. The mixing parameter μ measures the ratio of the external degree of a node with respect to its community to the total degree of the node.

The results of different methods on the GN benchmark networks are shown in Fig. 3. As can be seen, Louvain method performs fairly well on the GN benchmark network. This indicates that the community size of the GN benchmark network is not below the resolution limit, and the optimization of modularity indeed reveals the true partitions. LPAf performs next to Louvain method, and significantly better than the rest four methods. All the methods except Louvain and FineTune arrive at the same stable value of L at high μ , which corresponds to the trivial partition. LPAf cannot detect the real communities in this range by minimizing L , because the trivial partition has a lower L than the real partition.

We also adopted the LFR benchmark [28], which is a special case of the planted l -partition model [40]. LFR

networks are similar to real-world networks, since all of them are characterized by heterogeneous distributions of node degrees and community sizes. In our experiments, the parameters are fixed as follows: node degrees and community sizes are governed by the power law, with exponents being -2 and -1 respectively; the maximum degree is 50; the ranges of community sizes are [10,50] and [20,100] for smaller and bigger communities respectively; the network size N is either 1000 or 5000. The significance of community structure is controlled by a mixing parameter $\mu \in [0, 1]$ where smaller values correspond to more obvious community structure. μ is the expected fraction of links of a node connecting to other communities.

Results are assessed in terms of average NMI, shown in Fig. 4, which shows that, the LPAf outperforms other methods consistently for a wide range of μ . In contrast to the GN benchmark, Louvain method fails to detect the real communities even when μ is small for larger networks with smaller communities. This is due to the well-known resolution limit of modularity, i.e., there exists a size cutoff below which modularity cannot identify communities [26]. In order to optimize modularity, Louvain method tends to merge natural communities into much larger ones, which leads to rather poor performance. FineTune does not have remarkable performance either, as it also starts to fail for low values of μ . The nsdLPA performs better than LPA due to the consideration of the positive neighbor strength. Infomap performs comparably with LPAf in larger networks but is outperformed by LPAf when networks are small. Moreover, LPAf is more stable than LPA, because of the lower standard deviation of its NMI scores. The results thus confirm that LPAf performs better than or at least as well as the rest five methods in all the LFR networks.

To further address the validity of LPAf, we also computed the average ratio of the number of detected communities to the number of actual ones and showed them

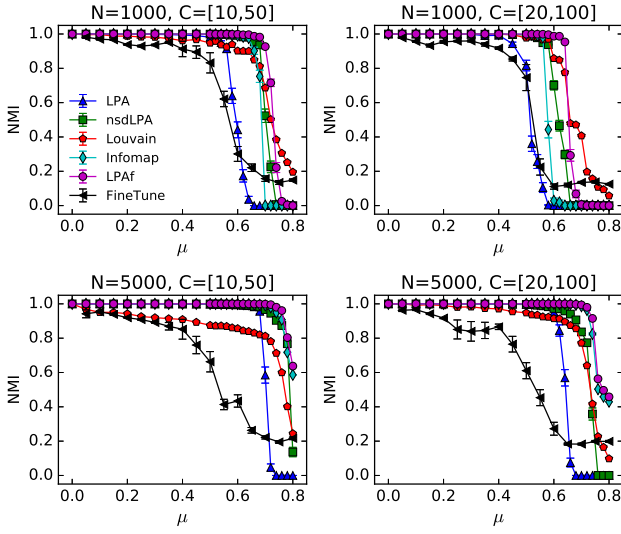


Fig. 4. (Color online) Average NMI for different algorithms as a function of μ on LFR networks. The number of vertices is either 1000 (small scale) or 5000 (large scale), and the ranges of community sizes are $[10, 50]$ (smaller community, left panel) and $[20, 100]$ (larger community, right panel). Each data point is averaged over 100 different network realizations.

in Fig. 5. As can be seen, the number of communities detected by the LPAf is very close to the actual one up to a high μ in all cases. The number of communities detected by nsdLPA is larger than the actual one at high values of μ , which implies that nsdLPA tends to form local subgroups and favors smaller communities due to the consideration of neighborhood strength. Louvain method tends to find less communities than planted ones due to the resolution limit of modularity, whereas FineTune normally detects more communities than actual ones in most cases. This indicates that FineTune resolves, to a certain degree, the resolution problem of Louvain method. In most cases, Infomap tends to find slightly less communities than actual ones.

To compare the computational loads of different methods, we plot the average elapsed times in Fig. 6. Generally, the running times of all methods increase when μ gets larger. This is due to that when μ is small, the communities are well separated and all the methods can easily de-

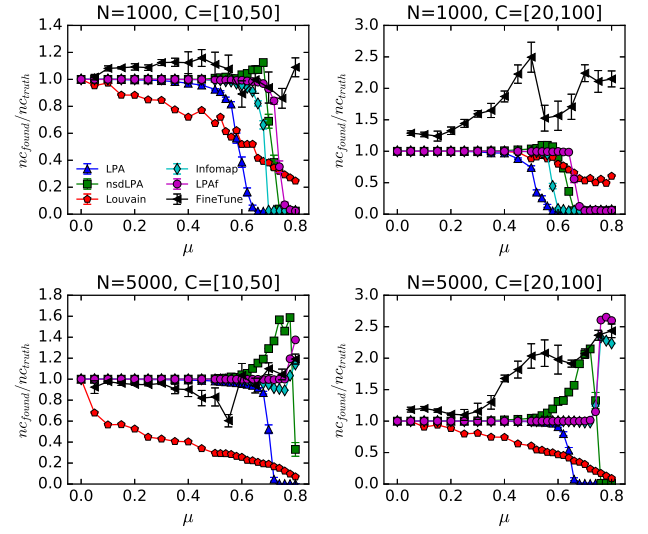


Fig. 5. (Color online) Average ratio of the number of detected communities to the number of actual communities for different algorithms with varying μ on LFR networks. Each data point is averaged over 100 different network realizations.

tect them in a short period of time. When μ increases to a specific value where the community structure still persists but is much more difficult to be revealed, the convergence speed slows down and thus results in peaks of the curves. When μ continues increasing, most of the methods cannot detect non-trivial communities and converge slightly faster than at the transition stage. Specifically, LPA and nsdLPA are faster than the rest four algorithms. LPAf, Louvain and Infomap exhibit similar time consuming patterns.

To test how well LPAf performs in finding the local minimum in L space, we computed the values of L for the partitions detected by LPAf and plotted them in Fig. 7. Due to the global minimum of L is not available, L -values of the planted partitions are adopted as references. As can be seen, when μ is small, i.e., the community structure is clear enough, the detected partitions and the true partitions almost have the same values of L , which indicates that LPAf correctly finds the real communities in the corresponding range of μ . When μ increases to a specific

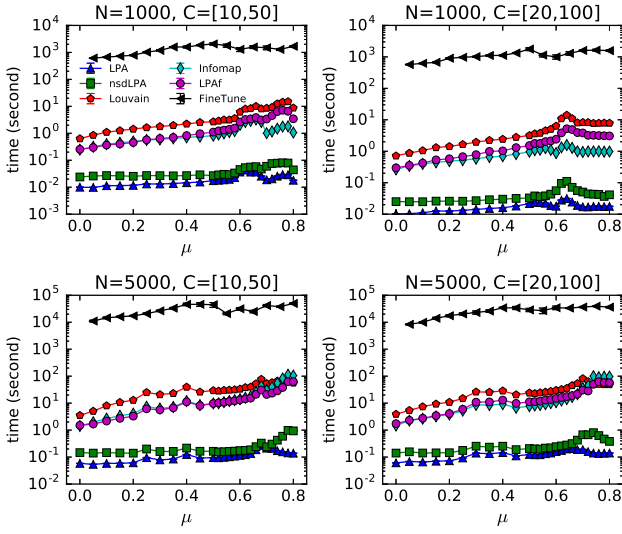


Fig. 6. (Color online) Average elapsed time as a function of μ of different algorithms on the LFR networks. Each data point is an average over 100 different network realizations.

value, L decreases rapidly to a stable value on smaller networks, which corresponds to the trivial partition that the whole network is regarded as a single community. As the trivial partition has a lower value of L than the planted one above a certain value of μ , LPAf cannot detect any non-trivial communities within this range. However in larger networks, LPAf yields larger L than that of the planted partition above a certain value of μ , which implies that LPAf is trapped in a suboptimal valley in L space.

3.2 Tests on real-world networks

We also applied the algorithms to several real-world networks that are commonly used for tests. The details of such networks are listed in Table 1.

We first compared directly the stability of different methods. All the methods are applied to each network 1000 times and the numbers of distinct detected partitions are reported. The pairwise VOI of the partitions are also computed to further evaluate the robustness of the methods. FineTune is not considered here since it is a deterministic algorithm. Due to the time complexity, two

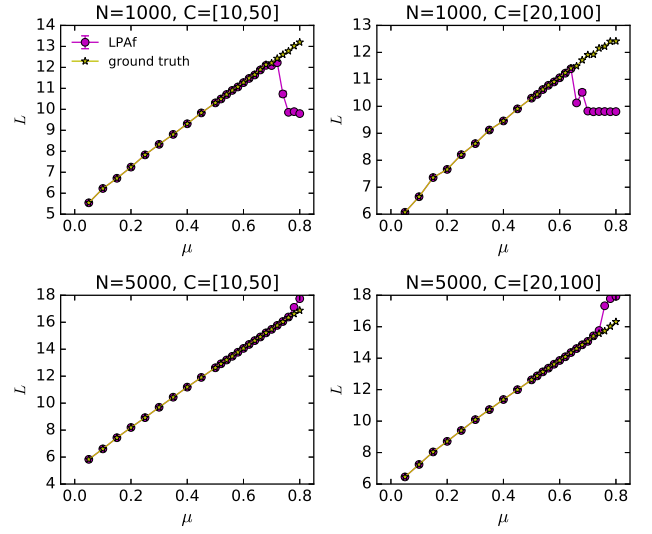


Fig. 7. (Color online) The comparison of L -values between detected and real partitions.

Table 1. Real-world networks with community structure.

Network	Reference	Vertices	Edges
karate	Zachary's karate club [41]	34	78
dolphins	Dolphin social network [42]	62	159
books	Books about US politics [43]	105	441
football	American College football [44]	115	613
blogs	Political blogs [45]	1490	16715
netsci	Network scientists [6]	1589	2742
power	US power grid [46]	4941	6594
mat-cond	Condensed matter collaborations [1]	16726	47594

larger networks, *power* and *mat-cond*, are excluded from the analysis. Results are shown in Tables 2 and 3. It is shown that LPAf is comparatively stable with less distinct partitions in most cases. LPA and Infomap are relatively unstable, even on smaller networks. Louvain method and nsdLPA have similar robustness, except on the *netsci* network where Louvain method yields the most stable results. Moreover, as shown in Table 3, the values of pairwise VOI between the partitions revealed by LPAf are lower than those for other methods in most cases. This concludes that LPAf is significantly more robust than LPA, and performs fairly stable.

Table 2. Analysis of the stability of different methods. We report the number of distinct community structures obtained over 1000 runs.

Network	LPA	nsdLPA	LPAf	Louvain	Infomap
karate	81	11	9	23	32
dolphins	425	52	72	39	609
books	191	75	10	73	725
football	464	78	33	47	706
netsci	1000	1000	496	181	1000

Table 3. Analysis of the stability of different methods. We report the average pairwise VOI of the corresponding partitions obtained over 1000 runs.

Network	LPA	nsdLPA	LPAf	Louvain	Infomap
karate	0.5189(4)	0.2269(2)	0.00482(2)	0.1967(2)	0.2021(2)
dolphins	0.4308(2)	0.1387(1)	0.2130(1)	0.2089(2)	0.3177(1)
books	0.2989(1)	0.1803(1)	0.03818(9)	0.1677(1)	0.3095(1)
football	0.14251(9)	0.05192(4)	0.02157(4)	0.05211(6)	0.12204(8)
netsci	0.037384(6)	0.027995(5)	0.008604(5)	0.006858(5)	0.018963(6)

Next, we detailedly analyzed the three networks (*karate*, *dolphins*, and *football*) which have known community structures. Fig. 8 shows the communities detected by LPAf on *karate* and *dolphin* networks with the lowest L . Zachary’s *karate* club is a social network of friendships between 34 members of a karate club at a US university in the 1970s. It splits into two smaller clubs after a dispute between club president John (node 34) and instructor Mr. Hi (node 1). As can be seen, three communities are discovered in this network by our algorithm. One of the two real communities is divided into two small ones (as shown in Fig. 8 (left panel)). The dolphin social network describes the frequent associations between 62 dolphins living off Doubtful Sound, New Zealand. The links represent that dolphins are observed to stay together more often than expected by chance during the years from 1994 to 2001. Four communities identified by our algorithm in this network are shown in Fig. 8 (right panel).

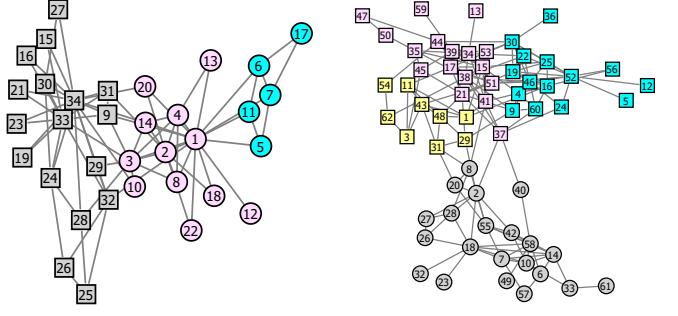


Fig. 8. (Color online) Communities detected by LPAf on *karate* (left panel) and *dolphin* (right panel) networks. The detected communities are distinguished by their colors, whereas the actual communities are represented by node shapes.

The *football* network describes football games among Division IA colleges during regular season Fall 2000. As shown in Fig. 9, the 115 nodes in the network represent teams, which are grouped into eleven different conferences, except for five independent teams. The regular season games between each pair of teams are shown as 613 edges of the network. Our algorithm identifies eleven communities within this network, as shown in Fig. 9. Among them, eight conferences are correctly identified. The three remaining communities closely resemble the Conference USA, Sun Belt and Western Athletic conferences. Five independent teams that do not belong to any conference tend to be grouped with the conferences which they are most closely associated.

For comparison, we applied different methods to *karate*, *dolphins*, *books* and *football* networks, and measured the NMI between the real partitions and those detected by different methods. The average values of NMI over 1000 runs are shown in Table 4. FineTune is deterministic and thus we only run it once. As one can see, LPAf performs fairly well on the *karate* and the *football* networks, although not the best. However it does not work well on the other two networks. The reason could be that the known partitions of these two networks do not have the lowest values of L , which prevents LPAf from detecting the real communities on these networks.

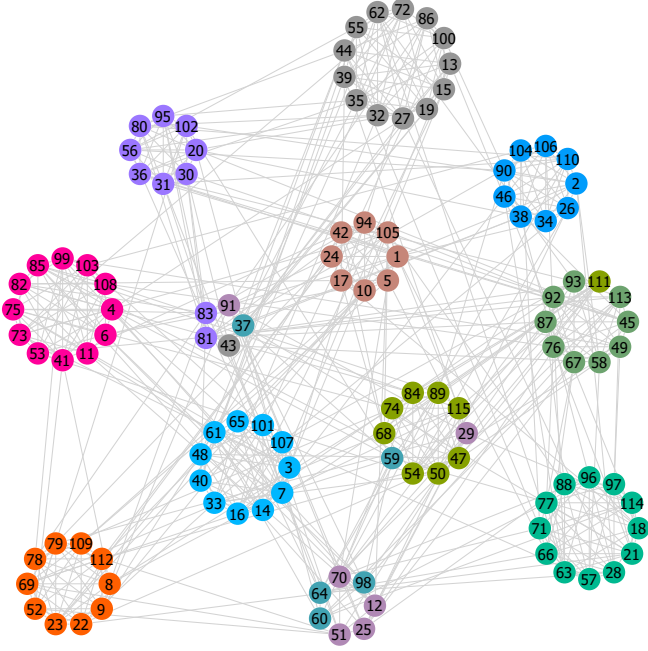


Fig. 9. (Color online) The *football* network with each node representing a NCAA team, and each edge denoting a game played in 2000 between two teams they co-join. Colors represent the communities detected by LPAf. The 12 NCAA conferences are grouped into circles.

Table 4. Average NMI between the real communities and those identified by the algorithms. Results are averages over 1000 different runs.

Network	LPA	nsdLPA	LPAf	Infomap	Louvain	FineTune
karate	0.689(7)	0.833(3)	0.821(1)	0.751(2)	0.651(1)	0.5925
dolphins	0.622(3)	0.606(1)	0.520(1)	0.506(1)	0.493(1)	0.4338
books	0.5494(9)	0.5395(7)	0.5391(3)	0.5414(9)	0.5421(7)	0.4146
football	0.8834(9)	0.9039(3)	0.9197(3)	0.8994(6)	0.8787(5)	0.9242

In Table 5, we also reported average modularity Q of the detected partitions for all networks so as to enable a complete comparison. It is not surprising that Louvain method yields the highest values on almost all networks, because it is based on the optimization of modularity. Therefore, for clarity, we show the results of Louvain method in the rightmost column of the table. We also mark the best results of the rest methods in bold type. As one can see, LPAf achieves the best performance among

Table 5. Average modularity Q of partitions identified by different algorithms. Results are the averages over 1000 different runs. Except the Louvain method, the best values are marked as boldface.

Network	True	LPA	nsdLPA	LPAf	Infomap	FineTune	Louvain
karate	0.3718	0.344(3)	0.3747(2)	0.4008(1)	0.3994(2)	0.4174	0.4154(2)
dolphins	0.3787	0.482(1)	0.5239(1)	0.5216(2)	0.5067(5)	0.4547	0.5206(1)
books	0.4149	0.4959(5)	0.5183(2)	0.52641(6)	0.5163(2)	0.4855	0.52626(6)
football	0.554	0.5893(4)	0.5673(5)	0.60052(4)	0.5907(2)	0.6005	0.60402(5)
netsci		0.9028(1)	0.9093(1)	0.9314(1)	0.9313(2)	0.7641	0.95904(2)
power		0.7175(4)	0.7204(3)	0.8295(2)	0.8297(2)	0.6036	0.93584(7)
mat-cond		0.7167(3)	0.7270(2)	0.7695(1)	0.7758(2)	0	0.8479(1)

Table 6. Average modularity density Q_{ds} of partitions identified by different algorithms. Results are the averages over 1000 different runs. Except the FineTune method, the best values are marked as boldface.

Network	True	LPA	nsdLPA	LPAf	Infomap	Louvain	FineTune
karate	0.1823	0.197(2)	0.1849(7)	0.2168(1)	0.2164(8)	0.2284(3)	0.231
dolphins	0.1362	0.184(2)	0.1967(7)	0.2060(5)	0.196(1)	0.2009(9)	0.264
books	0.1267	0.174(1)	0.1952(7)	0.1986(3)	0.193(1)	0.1972(4)	0.2506
football	0.4281	0.432(3)	0.465(2)	0.482(2)	0.457(2)	0.437(2)	0.4909
netsci		0.6417(6)	0.6409(4)	0.6136(4)	0.6093(5)	0.5029(3)	0.4866
power		0.2309(3)	0.2339(3)	0.1527(2)	0.1462(2)	0.02067(7)	0.3106
mat-cond		0.3036(3)	0.2979(3)	0.2526(1)	0.2401(2)	0.07047(9)	0.0003

those methods which do not directly optimize modularity in most cases.

In Table 6, we presented average modularity density Q_{ds} of partitions detected by different methods. FineTune is based on the optimization of modularity density. Therefore, we show the results of FineTune in the last column of the table and highlight the best results of the rest methods in bold type. As one can see, LPAf performs quite well in terms of Q_{ds} in most cases.

In Table 7, we compared different methods in terms of L . L -values of the true partitions are presented as references. As seen, LPAf achieves the best performance in most cases. It should be pointed out that the true partitions do not possess the global minimum of L . LPAf always obtains a lower L than that of the true partition in some networks. This explains why LPAf cannot detect the real communities correctly on these networks.

Table 7. Average description length L of partitions identified by different algorithms. Results are the averages over 1000 different runs.

Network	True	LPA	nsdLPA	LPAf	Infomap	Louvain	FineTune
karate	4.3408	4.392	4.3483	4.2996(4)	4.319(1)	4.418(1)	4.4018
dolphins	5.0786	5.068	4.9982(9)	5.099(1)	5.159(2)	5.095(1)	5.7197
books	6.0373	5.658(2)	5.618(2)	5.5875(7)	5.653(2)	5.603(1)	6.1893
football	6.3784	6.091(3)	6.311(4)	6.0503(3)	6.104(2)	5.9811(4)	6.055
netsci		4.135(1)	4.061(1)	3.7949(4)	3.8142(7)	3.9716(5)	6.7201
power		8.467(1)	8.417(1)	6.8032(6)	6.8549(6)	7.348(1)	10.4736
mat-cond		9.419(4)	9.239(3)	8.497(1)	8.608(2)	9.125(2)	13.4179

Lastly, we further analyzed the two larger networks, *power* and *mat-cond*. For simplicity, we only compared LPA and LPAf. We ran each method 100 times and analyze the conductances of the detected communities at various scales. The results are given in the form of NCP plots, as shown in Fig. 10. NCP plots evaluate the quality of the best community (in terms of conductance) as a function of its size. Previous studies show that many kinds of real-world networks exhibit a common characteristic structure of NCP plots, i.e., initial decreasing and subsequent increasing trend [34].

In the case of *power* network, LPAf detects communities on a much boarder scale with significant lower conductances, including also larger communities with around 80 nodes. On the *mat-cond* network, both LPAf and LPA find the best communities at the same scale (i.e, at around 15 nodes), while the conductances of LPAf are sightly lower than that of LPA. Note that LPA reveals a number of larger communities with significant high conductances in both networks (i.e., blue circles in the top right part of the top two plots of Fig. 10), which could be that many tie-breaks encountered in the label propagation process contributes to the formation of some large communities with high conductances.

3.3 Time complexity

Given a network with n nodes and m edges, let k be the maximum degree of nodes in this network. The time com-

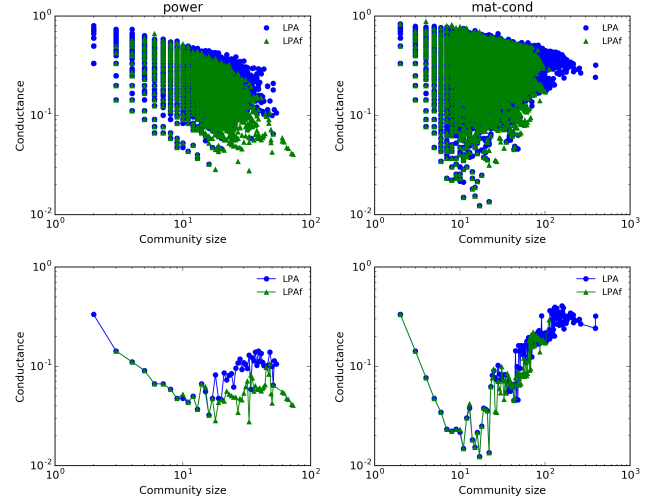


Fig. 10. (Color online) Comparison of LPAf and LPA on *power* and *mat-cond* networks. The conductances of individual communities (top panel), and the minimum conductances (bottom panel) at different scales are presented. Results were obtained over 100 runs.

plexity of each step of LPAf is roughly estimated as follows:

1. *Initialization* takes time of $O(n)$. Assigning a unique label to each node takes time of $O(n)$.
2. *Label propagation* takes time at most $O(nk)$. For each node, it iterates through at most k neighbors, thus, the upper bound of cost time of this step is $O(nk)$.
3. *Merging communities* takes time at most $O(m \log n)$. Merging pairs of communities using MSG requires a time of $O(m \log n)$ in the worst case (see Ref. [29] for detailed analysis).

Steps 2 and 3 are repeated, so the time per iteration is $O(kn + m \log n)$. Consequently, the time complexity of LPAf is roughly $O(kn + m \log n)$.

To evaluate the efficiency of LPAf, we have run LPA, nsdLPA, LPAf, Infomap and Louvain method on LFR networks with different sizes. Due to the high time complexity, FineTune is not considered in the benchmark situation. We repeated each experiment 30 times and reported the average running times. As shown in Fig. 11, the time

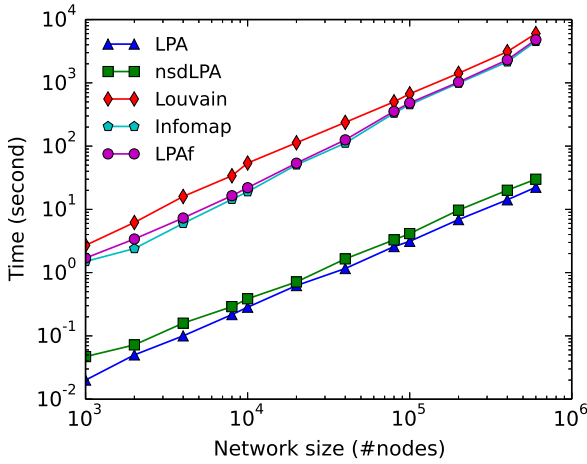


Fig. 11. (Color online) Comparison of running time for different algorithms on LFR networks with various sizes. Each data point is averaged over 30 different runs. The parameters of LFR networks are: the average degree is 20, the maximum degree is 50, the mixing parameter $\mu = 0.4$ and the range of community sizes $C = [20, 100]$.

complexity of LPA and nsdLPA is quite lower compared to the rest three methods. Still, all methods exhibit near linear time complexity and can be easily scaled to larger networks.

4 Conclusion

In this paper, we propose a modified label propagation algorithm (LPAf) to detect community structures in networks. In this algorithm, we introduce a new update rule which updates the label of a node by compressing a description of probability flow. Besides, by employing a multi-step greedy agglomerative algorithm, we merge pairs of communities so as to escape local minima in L -space. Furthermore, an incomplete update condition is adopted to accelerate the convergence.

We test the proposed algorithm on both synthetic and real-world networks, and compare its performance with that of the other five widely used methods in terms of *modularity*, *modularity density*, NMI, VOI and *conductance*. Firstly, we find that LPAf performs very well on

synthetic networks. In contrast to the Louvain method, LPAf is able to detect small communities in large networks. Secondly, we find that, LPAf detects communities which have lower conductances than that of LPA; by minimizing L , LPAf may fail to detect the real community structure which does not have the lowest L ; LPAf is generally more stable than LPA. Finally, we analyze the time complexity of LPAf and find that it depends linearly on the network size in sparse networks.

In the future work, we intend to test our algorithm on weighted and directed networks. We also plan to extend our approach to overlapping community detection by allowing each node possess multi-labels.

Acknowledgements

This work was in part supported by the Program of Introducing Talents of Discipline to Universities under grant no. B08033, and National Natural Science Foundation of China (Grant No. 11505071).

Author contribution statement

J.H. designed the algorithm, implemented the experiments, and prepared all the figures. J.H., L.Z. and Z.S. analyzed the results. All authors wrote, reviewed and approved the manuscript.

Appendix A: The change of average description length when a node moves from one community to another

From Eq. (4), for undirected and unweighted networks, the change of average description length when a node α

updates its label from i to j is given by,

$$\begin{aligned}\Delta L(\alpha, i, j) = & (q_{i\sim} + \delta q_{i\sim}) \log(q_{i\sim} + \delta q_{i\sim}) - q_{i\sim} \log(q_{i\sim}) \\ & - 2[(q_{i\sim} + \delta q_{i\sim}) \log(q_{i\sim} + \delta q_{i\sim}) - q_{i\sim} \log(q_{i\sim})] \\ & - 2[(q_{j\sim} + \delta q_{j\sim}) \log(q_{j\sim} + \delta q_{j\sim}) - q_{j\sim} \log(q_{j\sim})] \\ & + (p_{i\cup} + \delta p_{i\cup}) \log(p_{i\cup} + \delta p_{i\cup}) - p_{i\cup} \log(p_{i\cup}) \\ & + (p_{j\cup} + \delta p_{j\cup}) \log(p_{j\cup} + \delta p_{j\cup}) - p_{j\cup} \log(p_{j\cup})\end{aligned}$$

with

$$\begin{aligned}\delta q_{i\sim} &= \delta q_{i\sim} + \delta q_{j\sim}, \\ \delta p_{i\cup} &= \delta q_{i\sim} - p_{\alpha}, \\ \delta p_{j\cup} &= \delta q_{j\sim} + p_{\alpha}, \\ \delta q_{i\sim} &= \sum_{\beta \in \partial\alpha \cap V_i} \frac{1}{2m} - \sum_{\beta \in \partial\alpha \setminus V_i} \frac{1}{2m}, \\ \delta q_{j\sim} &= \sum_{\beta \in \partial\alpha \setminus V_j} \frac{1}{2m} - \sum_{\beta \in \partial\alpha \cap V_j} \frac{1}{2m},\end{aligned}$$

where m is the total number of edges of the network, V_i and V_j are the nodes in community i and j respectively, and $\partial\alpha$ is the neighbors of α . Extension to directed and weighted networks is straightforward.

References

1. M.E.J. Newman, Proceedings of the National Academy of Sciences of the United States of America **98**, 404 (2001), 0007214
2. J. Scott, *Social Network Analysis: A Handbook*, Vol. 3 (2000), ISBN 0761963391, <http://www.amazon.com/dp/0761963391>
3. D.A. Fell, A. Wagner, Nat Biotech **18**, 1121 (2000)
4. S. Fortunato, Physics Reports **486**, 75 (2010)
5. M.E.J. Newman, M. Girvan, Phys. Rev. E **69**, 026113 (2004)
6. M.E.J. Newman, Phys. Rev. E **74**, 036104 (2006)
7. S. White, P. Smyth, *A Spectral Clustering Approach To Finding Communities in Graph.*, in *SDM* (2005), citeseer.ist.psu.edu/734075.html
8. J. Reichardt, S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004)
9. S.W. Son, H. Jeong, J.D. Noh, The European Physical Journal B - Condensed Matter and Complex Systems **50**, 431 (2006)
10. J.M. Kumpula, J. Saramki, K. Kaski, J. Kertsz, The European Physical Journal B **56**, 41 (2007)
11. H. Zhou, R. Lipowsky, in *Computational Science - ICCS 2004*, edited by M. Bubak, G. van Albada, P. Sloot, J. Dongarra (Springer Berlin Heidelberg, 2004), Vol. 3038 of *Lecture Notes in Computer Science*, pp. 1062–1069, ISBN 978-3-540-22116-6, http://dx.doi.org/10.1007/978-3-540-24688-6_137
12. P. Pons, M. Latapy, in *Computer and Information Sciences - ISCIS 2005*, edited by p. Yolum, T. Gngr, F. Grgen, C. zturan (Springer Berlin Heidelberg, 2005), Vol. 3733 of *Lecture Notes in Computer Science*, pp. 284–293, ISBN 978-3-540-29414-6, http://dx.doi.org/10.1007/11569596_31
13. Ochab, J.K., Burda, Z., Eur. Phys. J. Special Topics **216**, 73 (2013)
14. V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment **2008**, P10008 (2008)
15. Barber, Michael J., Eur. Phys. J. B **86**, 385 (2013)
16. Waltman, Ludo, Jan van Eck, Nees, Eur. Phys. J. B **86**, 471 (2013)
17. Xiang, J., Hu, X.G., Zhang, X.Y., Fan, J.F., Zeng, X.L., Fu, G.Y., Deng, K., Hu, K., Eur. Phys. J. B **85**, 352 (2012)
18. M.E.J. Newman, E.A. Leicht, Proceedings of the National Academy of Sciences **104**, 9564 (2007), <http://www.pnas.org/content/104/23/9564.full.pdf>
19. M. Mungan, J.J. Ramasco, Journal of Statistical Mechanics: Theory and Experiment **2010**, P04028 (2010)
20. W. Ren, G. Yan, X. Liao, L. Xiao, Phys. Rev. E **79**, 036111 (2009)

21. J.M. Hofman, C.H. Wiggins, *Phys. Rev. Lett.* **100**, 258701 (2008)
22. U.N. Raghavan, R. Albert, S. Kumara, *Phys. Rev. E* **76**, 036106 (2007)
23. G. Tibly, J. Kertsz, *Physica A: Statistical Mechanics and its Applications* **387**, 4982 (2008)
24. M.J. Barber, J.W. Clark, *Phys. Rev. E* **80**, 026129 (2009)
25. X. Liu, T. Murata, *Physica A: Statistical Mechanics and its Applications* **389**, 1493 (2010)
26. S. Fortunato, M. Barthlemy, *Proceedings of the National Academy of Sciences* **104**, 36 (2007), <http://www.pnas.org/content/104/1/36.full.pdf>
27. I.X.Y. Leung, P. Hui, P. Liò, J. Crowcroft, *Phys. Rev. E* **79**, 066107 (2009)
28. A. Lancichinetti, S. Fortunato, F. Radicchi, *Phys. Rev. E* **78**, 046110 (2008)
29. A. Oades, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **77**, 1 (2008), 0712.1163
30. M. Rosvall, D. Axelsson, C.T. Bergstrom, *The European Physical Journal Special Topics* **178**, 13 (2009)
31. J. Xie, B. Szymanski, *Community detection using a neighborhood strength driven Label Propagation Algorithm*, in *Network Science Workshop (NSW), 2011 IEEE* (2011), pp. 188–195
32. M. Chen, T. Nguyen, B.K. Szymanski, *On Measuring the Quality of a Network Community Structure*, in *Social Computing (SocialCom), 2013 International Conference on* (2013), pp. 122–127
33. B. Bollobas, *Modern Graph Theory*, Graduate Texts in Mathematics (Springer New York, 1998), ISBN 9780387984889, <https://books.google.ca/books?id=SbZKSZ-1qrwC>
34. J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, *Internet Mathematics* **6**, 29 (2009)
35. D.J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003)
36. L. Danon, A. Daz-Guilera, J. Duch, A. Arenas, *Journal of Statistical Mechanics: Theory and Experiment* **2005**, P09008 (2005)
37. M. Meil, *Journal of Multivariate Analysis* **98**, 873 (2007)
38. M. Chen, K. Kuzmin, B.K. Szymanski, *IEEE Transactions on Computational Social Systems* **1**, 46 (2014)
39. M. Girvan, M.E.J. Newman, *Proceedings of the National Academy of Sciences* **99**, 7821 (2002), <http://www.pnas.org/content/99/12/7821.full.pdf>
40. A. Condon, R.M. Karp, *Random Structures & Algorithms* **18**, 116 (2001)
41. W.W. Zachary, *Journal of Anthropological Research* **33**, 452 (1977)
42. D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, S. Dawson, *Behavioral Ecology and Sociobiology* **54**, 396 (2003)
43. V. Krebs (2008)
44. M.E.J. Newman, *SIAM Review* **45**, 167 (2003), <http://dx.doi.org/10.1137/S003614450342480>
45. L.A. Adamic, N. Glance, *The Political Blogosphere and the 2004 U.S. Election: Divided They Blog*, in *Proceedings of the 3rd International Workshop on Link Discovery* (ACM, New York, NY, USA, 2005), LinkKDD '05, pp. 36–43, ISBN 1-59593-215-1, <http://doi.acm.org/10.1145/1134271.1134277>
46. D.J. Watts, S.H. Strogatz, *Nature* **393**, 440 (1998)